MAXIMUM LIKELIHOOD ESTIMATION IN A MULTINOMIAL MIXTURE MODEL

By

Charles E. McCulloch
Cornell University, Ithaca, N. Y.

BU-934-MA May, 1987

ABSTRACT

Maximum likelihood estimation is evaluated for a multinomial distribution, where the probabilities for each class are a linear combination of the unknown parameters. This model arises in genetic studies of multiple parentage.

I. INTRODUCTION

For many species of animals and insects, the ability of biologists to quantify multiple parentage within broods, clutches, litters, etc., is an important part of measuring reproductive success. For example, Dickinson (1986) estimated the proportion of offspring fathered by each of a female's two consecutive mates in a study of the milkweed leaf beetle. Often multiple parentage cannot be assessed from behavioral data alone (Sherman, 1981), but with the advent of starch-gel electrophoresis, parentage can sometimes be unambiguously assigned from genetic information. The purpose of this article is to indicate how maximum likelihood estimation can be used to estimate reproductive success in ambiguous cases and to evaluate the performance of the maximum likelihood estimator.

The small example that follows serves to illustrate the estimation problem. Consider a single locus example with two alleles (denoted 1 and 2) where there is a single mother and two possible fathers:

Male		
1	2	Female
11	12	12

.

Possible genotypes for the offspring are 11, 12 and 22. A genotype of 22 can be attributed unambiguously to male 2 but the others are ambiguous. If θ_1 denotes the probability of male 1 siring an offspring (and $\theta_2 = 1 - \theta_1$) and Mendelian assortment is assumed, the probabilities of the three genotypes are:

Genotype	Probability
11	$.5\theta_1 + .25\theta_2$
12	$.5\theta_1 + .5\theta_2$
22	$.25\theta_2$

This article focusses on estimating the probabilities, θ_i, of males siring offspring from data consisting of the genotypes of the offspring in a potentially multiply parented litter.

In the next Section we establish notation and define the basic estimation problem. Section 3 contains theoretical results about the estimation problem and the estimators and the results of a simulation study are reported in Section 4.

II. NOTATION AND BASIC RESULTS

For motivation sake, the context of the genetics example in the introduction will be retained. The data will consist of frequencies of occurrence of G genotypes, $f_1, f_2, \ldots, f_G$, which are assumed to be multinomially distributed with parameters

$$N = \sum_{i=1}^{G} f_i \quad \text{and} \quad \mathbf{p} = (p_1, p_2, \ldots, p_G),$$

where p_j = probability of the jth genotype. This implicitly assumes that individual offspring represent independent observations. This will likely be a good assumption under experimental conditions when known multiple mating has occurred. Under field conditions this may be a poor assumption.

Using the law of total probability, p_j can be expanded as follows:

$$p_j = \sum_{i=1}^{S} \theta_i p^*_{j|i},$$

where θ_i = probability of the ith male siring an offspring ($\sum_{i=1}^{S} \theta_i = 1$),

$p^*_{j|i}$ = conditional probability of genotype j given that the ith male is the sire,

and

S = total number of suspected sires.

The $p^*_{j|i}$ are assumed to be known (perhaps by calculation assuming Mendelian assortment and/or random mating) and the main interest is in estimating $\theta_1, \theta_2, \ldots, \theta_S$. Maximum likelihood can be used to estimate $\boldsymbol{\theta}$ and, since the expected values of the f_j are linear functions of the θ_i, ordinary least squares can also be used.

Using the notation $\mathbf{f} = (f_1, f_2, \ldots, f_G)'$, $\hat{\mathbf{p}} = \mathbf{f}/N$, and $\mathbf{P}^* = (p^*_{j|i})$, we have

$$E[\hat{\mathbf{p}}] = \mathbf{P}^{*\prime}\boldsymbol{\theta} \quad .$$

This suggests the ordinary least squares estimator

$$\hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{P}^* \mathbf{P}^{*\prime})^{-1} \mathbf{P}^* \hat{\mathbf{p}} \ ,$$

at least when $G \geq S$. As we will see later, the restriction $G \geq S$ is necessary. $\hat{\boldsymbol{\theta}}_{OLS}$ can be improved by requiring $\sum_{i=1}^{S} \hat{\theta}_i = 1$. This restricted, ordinary least squares estimator is still unbiased but has smaller variance than $\hat{\boldsymbol{\theta}}_{OLS}$. It is given by (Searle, 1977, p. 85)

$$\hat{\boldsymbol{\theta}}_{ROLS} = \hat{\boldsymbol{\theta}}_{OLS} + \frac{(\mathbf{P}^* \mathbf{P}^{*\prime})^{-1}\mathbf{1}}{\mathbf{1}'(\mathbf{P}^* \mathbf{P}^{*\prime})^{-1}\mathbf{1}} (1 - \mathbf{1}'\hat{\boldsymbol{\theta}}_{OLS}) \ , \qquad (2.1)$$

where $\mathbf{1}$ is a $S \times 1$ vector of all ones.

III. MAXIMUM LIKELIHOOD ESTIMATION

As noted in Section II, either maximum likelihood estimation or restricted, ordinary least squares are candidates for methods for estimating $\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}_{ROLS}$ will be unbiased, while the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{ML}$, will not be (see Example 2). $\hat{\boldsymbol{\theta}}_{ML}$ has the advantage that $0 \leq \hat{\boldsymbol{\theta}}_{ML} \leq 1$, while $\hat{\boldsymbol{\theta}}_{ROLS}$ can be outside the interval [0,1]. In terms of

calculation, $\hat{\theta}_{ROLS}$ may be relatively straightforwardly calculated while $\hat{\theta}_{ML}$ requires an iterative technique. We next discuss the calculation of $\hat{\theta}_{ML}$.

Using the multinomial assumption, the logarithm of the likelihood is given by

$$\log L = \sum_{j=1}^{G} f_j \log \left(\sum_{i=1}^{S} \theta_i p^*_{j|i} \right) \tag{3.1}$$

and

$$\frac{\partial^2 \log L}{\partial \theta_k \partial \theta_{k'}} = - \sum_{j=1}^{G} f_j \frac{p^*_{j|k}\, p^*_{j|k}}{p_j^{\,2}} \quad .$$

Writing $\mathbf{p}^*_j = (p^*_{j|1}, p^*_{j|2}, \ldots, p^*_{j|S})'$, the Hessian is given by

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{j=1}^{G} - \frac{f_j}{p_j^{\,2}} \mathbf{p}^*_j \mathbf{p}^{*\prime}_j$$

This is clearly a sum of negative semi-definite matrices and is therefore negative semi-definite. The log likelihood is therefore concave. To find the information matrix we first write $\theta_S = 1 - \sum_{i=1}^{S-1} \theta_j$ to introduce the restriction that $\sum_{i=1}^{S} \theta_i = 1$. This gives

$$\log L = \sum_{j=1}^{G} f_j \log \left(\sum_{i=1}^{S-1} \theta_i p^*_{j|i} + (1 - \sum_{i=1}^{S-1} \theta_i) p^*_{j|S} \right)$$

$$= \sum_{j=1}^{G} f_j \log \left(\sum_{i=1}^{S-1} \theta_i (p^*_{j|i} - p^*_{j|s}) + p^*_{j|S} \right)$$

and

$$\frac{\partial^2 \log L}{\partial \theta_k \partial \theta_{k'}} = \sum_{j=1}^{G} - \frac{f_j (p^*_{j|k} - p^*_{j|S})(p^*_{j|k'} - p^*_{j|S})}{p_j^2} \quad .$$

Therefore the information matrix is given by

$$I(\boldsymbol{\theta}) = E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = \left(\sum_{j=1}^{G} - \frac{(p^*_{j|k} - p^*_{j|S})(p^*_{j|k'} - p^*_{j|S})}{p_j} \right)_{kk'}.$$

Many numerical routines are available to maximize nonlinear functions such as the log likelihood given in (3.1), however, it is problematic to enforce the constraints on the parameters, namely

$$\begin{aligned} & 0 \le \theta_i \le 1 \\ & \sum_{i=i}^{S} \theta_j = 1 \\ & 0 \le p_j = \sum_{i=1}^{S} \theta_i p^*_{j|i} \le 1 \end{aligned} \tag{3.2}$$

while attempting to maximize the likelihood. A method that avoids these problems is the EM-algorithm. If the actual numbers of each genotype from each suspected sire were known, it would be very simple to form the maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_{ML}$, by using the proportion of offspring from each suspect. These can be estimated as "missing data" in the EM algorithm. Explicitly the algorithm is as follows, where k is the iteration, a superscript (k) indicates values at the kth iteration, $f_{i,j}$ denotes the part of f_j estimated to be from sire i and $p_{i|j}$ denotes the conditional probability of suspect i being the sire of genotype j.

0. $k = 0,\ \theta_i^{(k)} = \frac{1}{S}$

1. $k = k + 1,\ p_{i|j}^{(k)} = \dfrac{p^*_{j|i}\ \theta_i^{(k-1)}}{\sum_{r=1}^{S} p^*_{j|r}\ \theta_r^{(k-1)}}$

2. $f_{i,j}^{(k)} = f_j p_{i|j}^{(k)}$

3. $\theta_i^{(k)} = \sum_{j=1}^{G} f_{i,j}^{(k)} \Big/ \left(\sum_{j=1}^{G} f_j \right)$

4. If $\max_i [|\theta_i^{(k)} - \theta_i^{(k-1)}|] > \varepsilon$, return to step 1, otherwise stop.

This algorithm is easily programmed and runs quickly on a microcomputer. Extensive experience shows that it reliably converges to $\hat{\boldsymbol{\theta}}_{ML}$.

Several results about maximum likelihood estimation for this problem can be easily proven.

Proposition 1: Values of $\boldsymbol{\theta}$ that maximize the likelihood are not unique if $S > G$.

Proof: The likelihood is given by the elements of $\mathbf{P}^{*\prime}\boldsymbol{\theta}$ raised to the powers given by $\mathbf{f}$ and multiplied together. Thus, any values of $\boldsymbol{\theta}$ that give the same values for $\mathbf{P}^{*\prime}\boldsymbol{\theta}$ will give the same value for the likelihood. If $S > G$, then $\mathbf{P}^{*\prime}$ (which is $G \times S$) will have rank less than S and the equations

$$\mathbf{P}^{*\prime}\boldsymbol{\theta} = \mathbf{z}$$

will, if they have any solution, have a multitude of solutions (Searle, 1986, p. 235). Thus if a value of $\boldsymbol{\theta}$ gives a $\mathbf{z}$ which maximizes the likelihood it is not unique.

Due to symmetry and the equal starting values in the implementation of the EM-algorithm we have the following.

Proposition 2: Using the EM-algorithm, any suspects with identical $p^*_{j|i}$ ($j = 1, 2, \ldots, G$) will have identical estimates of $\boldsymbol{\theta}$.

Proposition 3: If $S = G$, $\mathbf{P}^*$ is of full rank, and if $\hat{\boldsymbol{\theta}}_{ROLS}$ satisfies the constraints (3.2) then $\hat{\boldsymbol{\theta}}_{ROLS} = \hat{\boldsymbol{\theta}}_{ML} = \hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{P}^{*\prime})^{-1}\hat{\mathbf{p}}$, where $\hat{\mathbf{p}} = \mathbf{f}/N$.

Proof: Clearly, $\hat{\mathbf{p}}$ is the MLE for the likelihood $L = \prod_{j=1}^{G} p_j^{f_j}$.

If $S = G$ and $\mathbf{P}^*$ is of full rank then $\mathbf{p} = (p_1, p_2, \ldots, p_S)'$ and $\boldsymbol{\theta}$ are related by a one-to-one transformation:

$$\mathbf{p} = \mathbf{P}^{*\prime}\boldsymbol{\theta} ,$$

$$(\mathbf{P}^{*\prime})^{-1}\mathbf{p} = \boldsymbol{\theta} .$$

By the invariance of maximum likelihood estimators if $(\mathbf{P}^{*\prime})^{-1}\hat{\mathbf{p}} = \hat{\boldsymbol{\theta}}_{OLS}$ satisfies the constraints (3.2) then $\hat{\boldsymbol{\theta}}_{ML} = (\mathbf{P}^{*\prime})^{-1}\hat{\mathbf{p}}$. Also if $\mathbf{P}^{*\prime}$ is square

$$\mathbf{1}'\hat{\boldsymbol{\theta}}_{OLS} = \mathbf{1}'(\mathbf{P}^{*\prime})^{-1}\hat{\mathbf{p}} = \mathbf{1}'\mathbf{P}^{*\prime}(\mathbf{P}^{*\prime})^{-1}\hat{\mathbf{p}} = \mathbf{1}'\hat{\mathbf{p}} = 1$$

since $\mathbf{1}'\mathbf{P}^{*\prime} = \mathbf{1}'$. Therefore the last term in $\hat{\boldsymbol{\theta}}_{ROLS}$ in equation (2.1) is zero and $\hat{\boldsymbol{\theta}}_{ROLS} = \hat{\boldsymbol{\theta}}_{OLS}$.

The following examples serve to show that Proposition 3 is not true if the conditions are not met. Below is an example where $S < G$.

Example 1: For this example $S = 2$, $G = 3$,

$$\mathbf{P}^{*\prime} = \begin{pmatrix} .5 & .875 \\ .25 & .125 \\ .25 & 0 \end{pmatrix},$$

and $\mathbf{f} = (8, 2, 0)'$. The log likelihood is given by

$$\log L = 8 \log(.875 - .375\theta_1) + 2 \log(.125 + .125\theta_1),$$

which is an increasing function of θ_1 up to $-1/3$ and then decreasing. Thus $\hat{\boldsymbol{\theta}}_{ML} = (0, 1)'$. $\hat{\boldsymbol{\theta}}_{ROLS}$ can be found using (2.1) which gives $\hat{\boldsymbol{\theta}}_{ROLS} = (6/35, 29/35)'$.

The following example illustrates that $\hat{\boldsymbol{\theta}}_{ML}$ is biased in general and that $\hat{\boldsymbol{\theta}}_{ROLS}$ is not restricted to the interval $[0, 1]$.

Example 2: Values of $\hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\theta}}_{ROLS}$ for the case where $S = 2$, $G = 2$, $\boldsymbol{\theta} = (1,0)$,

$$\mathbf{P}^{*} = \begin{pmatrix} .5 & .75 \\ .5 & .25 \end{pmatrix}$$

Frequencies f_1	Frequencies f_2	$P(F_1=f_1, F_2=f_2)$	Value of $\hat{\theta}_{ML,1}$	Value of $\hat{\theta}_{ROLS,1}$
10	0	.001	0	-1
9	1	.010	0	-.6
8	2	.044	0	-.2
7	3	.117	.2	.2
6	4	.205	.6	.6
5	5	.246	1	1.0
4	6	.205	1	1.4
3	7	.117	1	1.8
2	8	.044	1	2.2
1	9	.010	1	2.6
0	10	.001	1	3.0

Summary:

$\hat{\theta}_{ML}$: bias = -.23 variance = .11 root mean square error = .4

$\hat{\theta}_{ROLS}$: bias = 0 variance = .40 root mean square error = .63

IV. A SIMULATION STUDY

In this section we report the results of simulations of $\hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\theta}}_{ROLS}$. All the simulations were run on an IBM PC-AT using the language GAUSS and its built-in random number generators. A number of parameter configurations were used. They are given in Table 1. Sets A through C were used to investigate the effect of increasing the sample size and sets S through Z represent realistic groupings for the population used in the study described in Dickinson (1986). They illustrate the effect of changing the true $\boldsymbol{\theta}$.

Figures 1, 2 and 3 illustrate the effect of sample size on, respectively, the bias, root mean square error and the standard deviation for parameter configuration B (see Table 1). As can be seen, the bias of $\hat{\boldsymbol{\theta}}_{ML}$ is considerable for small sample sizes, but the root mean square error is always smaller than $\hat{\boldsymbol{\theta}}_{ROLS}$.

Configuration B is "typical" in the sense that estimation was neither very bad nor good compared to the other configurations. For every configuration studied the root mean square error for $\hat{\boldsymbol{\theta}}_{ML}$ was smaller or equal to that of $\hat{\boldsymbol{\theta}}_{ROLS}$.

V. CONCLUSIONS

In choosing between $\hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\theta}}_{ROLS}$, $\hat{\boldsymbol{\theta}}_{ML}$ is clearly better over the range of parameter values studied according to the criteria of mean square error. $\hat{\boldsymbol{\theta}}_{ROLS}$ should only be used if unbiasedness is paramount. However, with small sample sizes, neither estimator performed well. Thus, neither method of estimation could be expected to give accurate estimates, though in some cases the probability of a correct ranking was fairly high.

VI. REFERENCES

Dickinson, J. L. 1986. Prolonged mating in the milkweed beetle, *Labidomera clivicollis* (Coleoptera: Chrysomelidae), a test of the "sperm-loading" hypothesis. *Behav. Ecol. Sociobiol.* **18**: 331-338.

Searle, S. R. 1986. *Matrix Algebra Useful for Statistics*. Wiley, New York.

Seber, G. A. F. 1977. *Linear Regression Analysis*. Wiley. New York.

Sherman, P. W. 1981. Electrophoresis and Avian Genealogical Analyses. *The Auk*, **98**(2): 419-422.

cm

TABLE 1: Parameter Configurations for Simulations

Simulation set	S	G	Number of replications for simulation	P*	θ and number of offspring (NOBS)
A	3	3	1000	$\begin{bmatrix} .5 & .25 & .25 \\ 0 & .75 & .25 \\ .875 & .125 & 0 \end{bmatrix}$	θ = (.6, .35, .05) NOBS = 4,10,25,50,100
B	2	3	1000	$\begin{bmatrix} .5 & .25 & .25 \\ .875 & .125 & 0 \end{bmatrix}$	θ = (.75, .25) NOBS = 4,10,25,50,100
C	3	4	1000	$\begin{bmatrix} .5 & .25 & .25 & 0 \\ 0 & .75 & .25 & 0 \\ 0 & .25 & .5 & .25 \end{bmatrix}$	θ = (.5, .3, .2) NOBS = 4,10,25,50,100
S	2	2	1000	$\begin{bmatrix} .5 & .5 \\ 1 & 0 \end{bmatrix}$	NOBS = 10; θ = (1,0) (.95,.05), (.9,.1), (.8,.2), (.7,.3), (.6,.4), (.5,.5), (.4,.6), (.3,.7), (.2,.8), (.1,.9), (.05,.95), (0,1)
T	2	3	1000	$\begin{bmatrix} .5 & .5 & 0 \\ .25 & .5 & .25 \end{bmatrix}$	NOBS = 10; θ same as S
U	2	3	1000	$\begin{bmatrix} .25 & .25 & .25 & .25 & .25 \\ .5 & .5 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & .5 \end{bmatrix}$	NOBS = 10; θ = (1,0,0), (.8,.2,0), (.8,0,.2), (.6,.4,0), (.6,.2,.2) (.6,0,.4), (.4,.6,0), (.4,.4,.2) (.4,.2,.4), (.4,0,.6), (.2,.8,0) (.2,.6,.2), (.2,.4,.4),(.2,.2,.6) (.2,0,.8), (0,1,0), (0,.8,.2) (0,.6,.4), (0,.4,.6), (0,.2,.8) (0,0,1), (.333,.333,.333).
V	3	5	1000	$\begin{bmatrix} .25 & .25 & .25 & .25 & 0 \\ .5 & .5 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & .5 \end{bmatrix}$	NOBS = 25, θ same as U
W	2	4	1000	$\begin{bmatrix} .25 & .25 & .25 & .25 \\ .5 & .5 & 0 & 0 \end{bmatrix}$	NOBS = 25; θ same as S
X	2	5	1000	$\begin{bmatrix} .25 & .25 & .25 & .25 & 0 \\ 0 & 0 & 0 & .5 & .5 \end{bmatrix}$	NOBS = 25; θ same as S
Y	2	3	1000	$\begin{bmatrix} .5 & .5 & 0 \\ .25 & .5 & .25 \end{bmatrix}$	NOBS = 25; θ same as S
Z	2	2	1000	$\begin{bmatrix} .5 & .5 \\ 1 & 0 \end{bmatrix}$	NOBS = 25; θ same as S

FIGURE 1: Bias of MLE and ROLSE

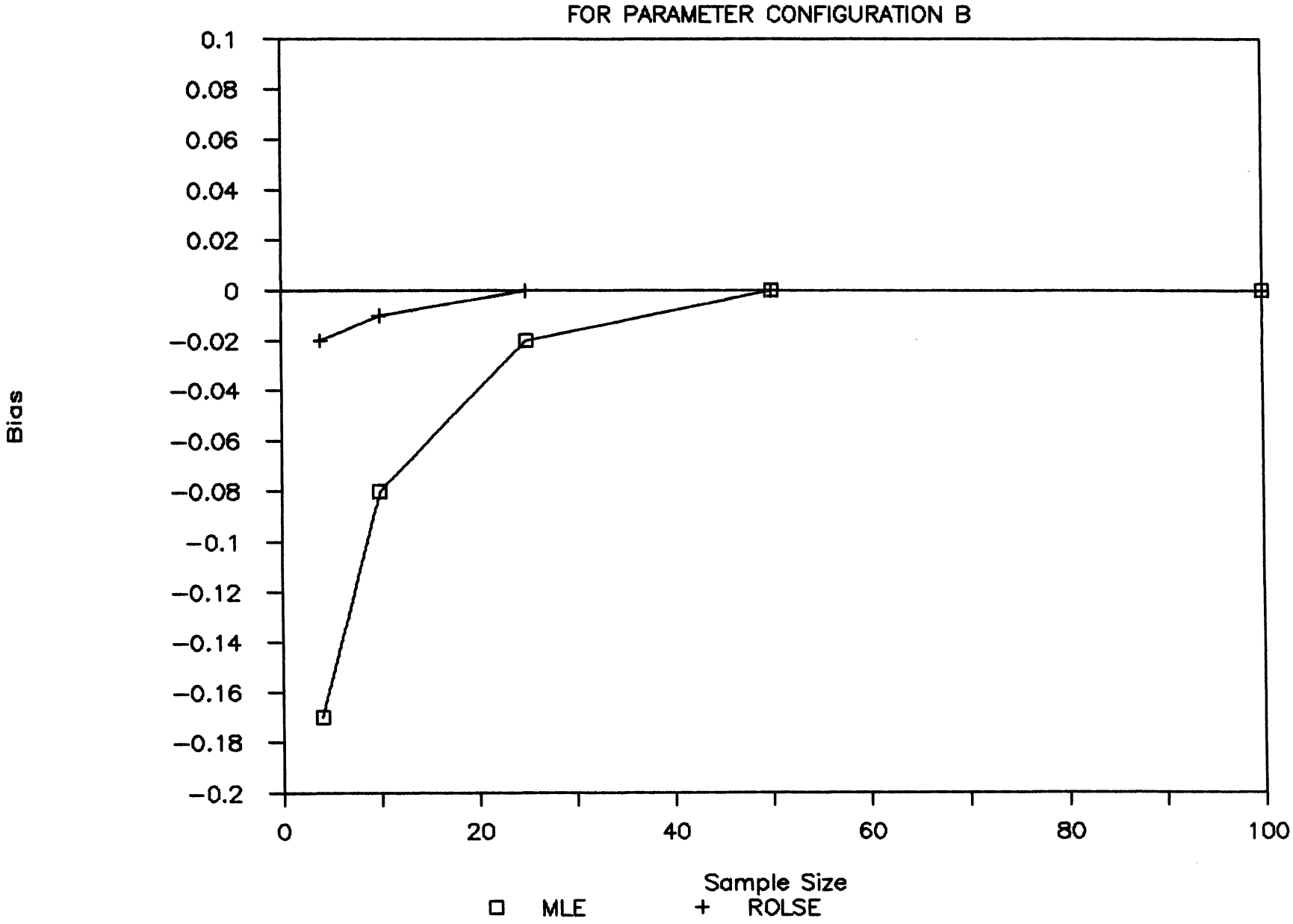

FIGURE 2: RMSE of MLE and ROLSE

FOR PARAMETER CONFIGURATION B

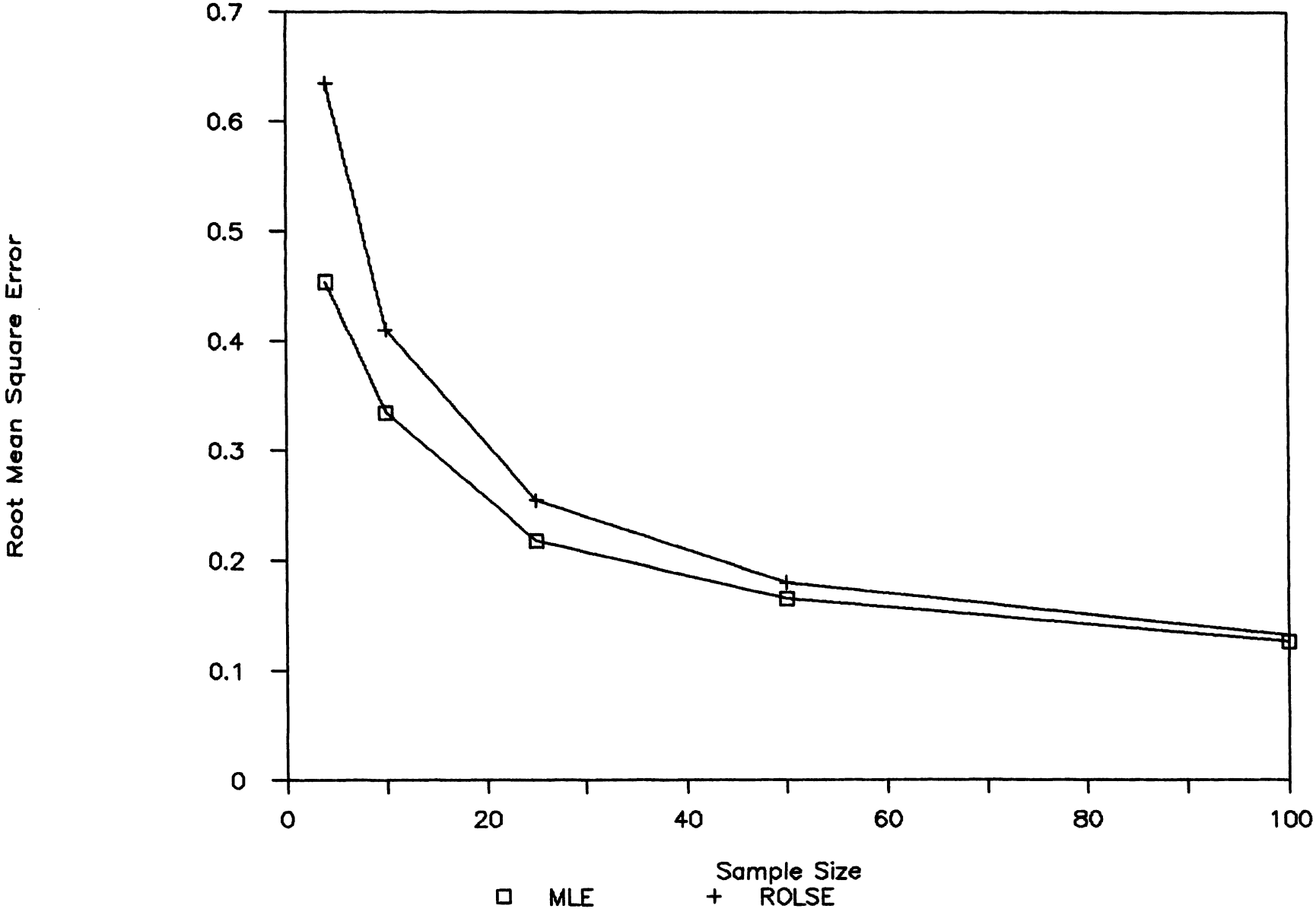

FIGURE 3: STDEV of MLE and ROLSE

FOR PARAMETER CONFIGURATION B